Generic Document Image Dewarping by Probabilistic Discretization of Vanishing Points

Gilles Simon, Salvatore Tabbone

Generic Document Image Dewarping by Probabilistic Discretization of Vanishing Points

Gilles Simon
Université de Lorraine, CNRS, Inria, LORIA
LORIA UMR 7503
Campus Scientifique, 615 Rue du Jardin-Botanique,
F-54506 Vandœuvre-lès-Nancy, France
Email: gilles.simon@loria.fr

Salvatore Tabbone
Université de Lorraine, CNRS, LORIA
LORIA UMR 7503
Campus Scientifique, 615 Rue du Jardin-Botanique,
F-54506 Vandœuvre-lès-Nancy, France
Email: antoine.tabbone@univ-lorraine.fr

***Abstract*—Document images dewarping is still a challenge especially when documents are captured with one camera in an uncontrolled environment. In this paper we propose a generic approach based on vanishing points (VP) to reconstruct the 3D shape of document pages. Unlike previous methods we do not need to segment the text included in the documents. Therefore, our approach is less sensitive to pre-processing and segmentation errors. The computation of the VPs is robust and relies on the *a-contrario* framework, which has only one parameter whose setting is based on probabilistic reasoning instead of experimental tuning. Thus, our method can be applied to any kind of document including text and non-text blocks and extended to other kind of images. Experimental results show that the proposed method is robust to a variety of distortions.**

I. Introduction

Optical character recognition (OCR) embedded into mobile cameras has become a big challenge nowadays [1]. Unlike scanned documents that are acquired with flatbed scanner with good lighting, document images captured by digital camera are subject to distortions such as perspective view or curved surfaces. Even with flatbed scanner for bounded documents non-linear warping are observed also. In this perspective, dewarping camera-captured document images have raised a lot of interest this last decades and many different approaches haven been proposed in the literature. The goal for dewarping algorithms is broadly to rectify a document page so that it is transformed into a flat one and that it appears in a frontal-flat view for an OCR algorithm. Several systems estimate the distortion following a 3-D shape reconstruction using extensions scanner hardware like stereo-cameras [2], [3], special light sectioning [4] or laser [5]. Even if these approaches are accurate to estimate the page surfaces their requirements of additional hardware limit their application areas since there are almost not portable. Therefore, most of the approaches focus on system with one camera in an uncontrolled environment. Some methods define 3D shape based on generalized cylindrical surface assumptions [6], [7], [8] or on texture flow information [9], [10]. Other approaches do not try to get a model of a 3D surface since they are based on 2D geometric features estimation[11], [12], [13]. These methods locally estimate a curve to line dewarping. Even if robust method has been proposed to segment documents[14], they are prone to errors of pre-processing steps (text line, word segmentation and binarization). Furthermore, as these methods mostly focus on text estimation distortions can appear for non-text regions (graphics and images). Moreover, it is difficult for these methods to recover foreshortening effects caused by perspective transformations. In parallel deep learning methods have been proposed[15], [16]. A CGAN network to learn page distortions is proposed in [17] but required an important number of learning samples which describe the space of all possible distortions. In this method perspective distortions are not considered. Even if 3D approaches are less sensitive to errors most of them rely on text line extraction.

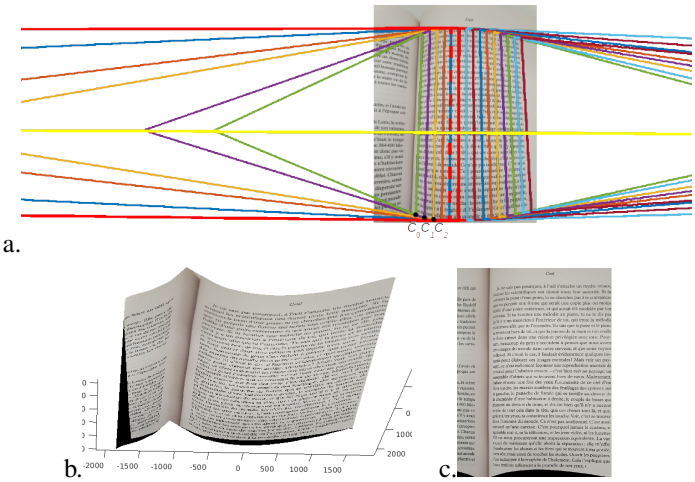


Fig. 1. Method overview. a. Vanishing point discretization. b. 3D reconstruction. c. Dewarped image.

Finally the page or open book can be traced by a straight line moving parallel to a fixed straight line (the vertical axis of the book), and intersecting a fixed horizontal curve, which is a particular case of cylindrical surface. Under this assumption, the problem of image rectification and curved lines (foreshortening effects) are jointly solved following the computation of the vanishing points (VPs) along the horizon line (HL). In particular, for curved surface patches, a continuous (possibly in pieces), one-dimensional locus of VPs (Sec. II-III) is obtained along the HL (Fig. 1.a). This locus is determined based on

a binary-tree descent searching of VPs, using a probabilistic criterion to decide when to stop subdivisions (Sec. IV). The camera's focal length is obtained from the zenith and the HL, which allows obtaining a 3D model of the document (Fig. 1.b), that just need to be "unfolded" to get the final dewarped image (Fig. 1.c) as described in Sec. V.

In [18], [19] two vanishing points were estimated following horizontal and vertical lines defined in the Radon domain. Perspective distortions are effectively removed for text documents but the approach is not suitable for geometric distortions due to curved pages as no reconstruction model was considered.

Our method is not based on any text line extraction or line tracing and does not rely on noise-sensitive operations such as image binarization and characters segmentation. The whole image is rectified which means that text and non-text distortions (image, graphic) are recovered. Overall, a small set of parameters easy to master need to be set and our approach relies on the probabilistic Line Segment Detector (LSD) [20] which is parameter free. The assumption of a cylinder surface is often verified in practice. Sometimes it is locally violated, but our method remains capable of properly dewarping the part of the page where it is verified. Experimental results on the well-known dataset IUPR are very effective and significant compared to the state-of-art (Sec. VI).

II. Horizon-first Vanishing Point Detection

Detection of VPs has recently made a qualitative leap forward in urban context, thanks to recent methods [21], [22], the main trick of which is to first determine the HL and only then calculate horizontal VPs along this line. Before this work, this sounded like a chicken-and-egg problem, because at least two horizontal VPs were required to compute the HL. Two different solutions have been proposed to get out of this dilemma, both based on the same framework which we also adopt. We now describe this framework, before indicating the limitations of the existing implementations with respect to the document dewarping problem.

A. General framework

The general framework is based on the fact that the HL is orthogonal to the zenith line, i.e. the line joining the principal point O (assumed at image centre in this work) to the zenith (Fig. 2.a). Thanks to this property, one can proceed as follows to detect the HL :

1) Compute the zenith line and the zenith.
2) Sample candidate HLs perpendicular to the zenith line.
3) Score each line according to the consistency of the two strongest VPs possibly detected on that line and choose the one with the highest score as the HL.
4) Calculate the VPs along the HL.

Step 1 is generally quite easy to solve, as many lines converge towards the zenith in images of buildings or documents. The zenith line is estimated differently in [21] and [22], but ultimately the zenith is calculated by singular value decomposition (SVD) using all line segments (LSs) parallel to a certain threshold to the estimated zenith line (implementation details can be found in the related papers).

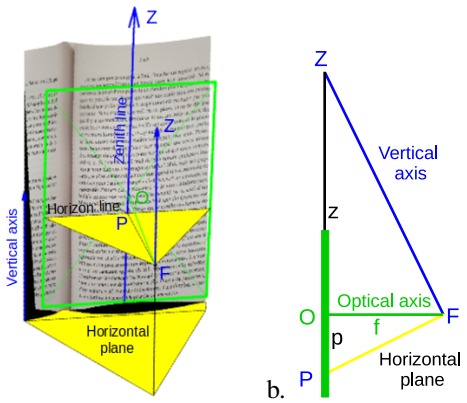


Fig. 2. Camera geometry. The horizon line is perpendicular to the zenith line (left), as is line (FZ) to line (FP) (left & right), where Z is the zenith VP, F is the camera's optical centre and P is the intersection point between the zenith line and the horizon line.

Step 3 is based on a calculation of possible VPs along each candidate line using the same method as for step 4. Both methods, as well as our own, use the same measure of angular consistency between a LS $\mathbf{l}_j$ and a VP $\mathbf{v}_i$:

$$f_c(\mathbf{v}_i, \mathbf{l}_j) = \max(\theta_{con} - |\sin^{-1}(\mathbf{v}_i^\top \mathbf{l}_j)|, 0), \quad (1)$$

with $\theta_{con} = 1.5°$. The two strongest VPs $\{\mathbf{v}_i\}_{best}$ are identified for each candidate line according to this measure, and the line's score is given by:

$$\sum_{\{\mathbf{v}_i\}_{best}} \sum_{\{\mathbf{l}_j\}} f_c(\mathbf{v}_i, \mathbf{l}_j). \quad (2)$$

Possibly the external sum includes only one VP or is equal to zero if no VP is detected on the line.

B. Limitation of existing implementations

The main difficulty with respect to work prior to [21] and [22] was to propose a way to calculate the offset probability density function (PDF) used for the sampling carried out in step 2, without any knowledge of VPs. The authors of [21] use a Gaussian model, fit from a categorical distribution generated by a convolutional neural network (CNN) for each HL parameter. The network was trained from tens of thousands of urban images accompanied by ground truth (GT) HLs. However, it has been shown in [22] that it is not relevant when used in a context other than the urban environment.

The authors of [22] rely on a probabilistic grouping of LSs detected by LSD. They noticed that horizontal LSs (in the scene) are projected parallel to and onto the HL when they are at the height of the optical center and non-parallel to that line when they are at a different height. This geometric property usually results in LSs (in the image) perpendicular to the zenith line accumulating at the height of the HL. A *a-contrario* model [23] is used to detect meaningful alignments of such LSs and the offset PDF is taken as a mixture of Gaussian, whose modes are the heights of the meaningful alignments, and the standard

deviations are set to $\sigma \times H$, where H is the image height and σ is set to 0.2 (Fig. 3.a-b, the PDF's mode is colored in cyan and the sampled lines in magenta). However, while this technique seems well suited to urban environments, it is not always suitable for document images, especially when parts of the text lines are near parallel to the HL. This can occur not only on flat surfaces, but also on curved surfaces, in vertical strips where the angle between the normal to the surface and the optical axis changes sign (e.g. the angular area shown by dashed and plain red lines in Fig. 1.a). In this case, almost any line of text inside the area can generate a PDF mode, which can cause inaccurate detection of the HL (yellow lines in Fig. 3.a-b, these lines are quite far from the predicted lines in cyan). For all these reasons, we adopt a different method to detect the HL, that is presented in Sec. III.

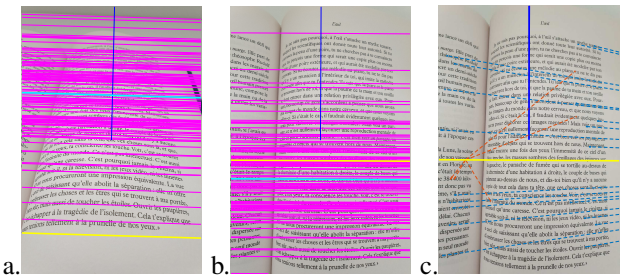


Fig. 3. Example results obtained by using the method of [22]. a-b. Horizon line sampling and detection. c. VP detection.

Finally, step 4 is probably the most critical in terms of dewarping. Both methods use LSs detected in the image to resolve this step. The approach of [21] consists in randomly selecting a subset of LSs $\{\mathbf{l}_j\}$ and computing their intersections with the horizon. An optimal subset of VPs $\mathbf{v}_i$ is extracted from the intersections, so that the sum of weights $\sum_{\mathbf{v}_i} \sum_{\mathbf{l}_j} f_c(\mathbf{v}_i, \mathbf{l}_j)$ is maximal, while ensuring no VPs in the final set are too close. However, for the page dewarping problem we need to detect very close VPs. In addition, this approach requires the use of several parameters that are difficult to adjust [22]. The method in [22] relies on an *a-contrario* detection of meaningful VPs along the HL (Sec. IV-A). As such it has no difficult parameters to set and obtains fewer spurious VPs than the previous approach. On the other hand, it is not suitable for curved surfaces on which only a few dominant VPs are usually detected (Fig. 3.c, one spurious and only one correct VPs are detected). By contrast, our method is capable of detecting hundreds of VPs on a curved surface (Sec. IV).

III. Computation of the Horizon Line

In both [21] and [22], the HL is detected in two stages: (i) a coarse prediction is obtained using a CNN or an *a-contrario* model; (ii) an arbitrary number of lines (300 with both methods) are sampled across the infinite image plane, using Gaussian PDFs centered on predictions. This procedure has the following drawbacks. First, if the prediction is incorrect, most samples are drawn close to the prediction and few in the area that actually contains the HL, which makes its eventual detection more imprecise (as in Fig. 3.a). On the other hand, although concentrated around the prediction, samples are potentially drawn in the entire image plane, which actually undercuts the pool of possible candidates. To tackle these issues, we propose to first delimit the sampling area based on some geometric constraints (Sec. III-A) and then to perform a prediction-free, coarse-to-fine sampling in this area (Sec. III-B).

A. Search for the zenith and the horizon line

The delimitation of the search areas for the zenith and the HL is based on two assumptions:

1) the page is roughly facing the camera. Specifically, the camera's tilt, relative to its orientation when facing the book, is assumed to be less than $45°$, as is its roll;
2) the focal length is in the range $[f_1, f_2]$, with $f_1 = 0.28W, f_2 = 3.8W$ where W is the image width.

These assumptions are quite broad. The first corresponds to fairly natural recommendations that must be respected to take the picture. The second corresponds to a wide range of focal lengths that cover a large number of cameras. It has also been used e.g. in [24] to reject miscalculated focal lengths.

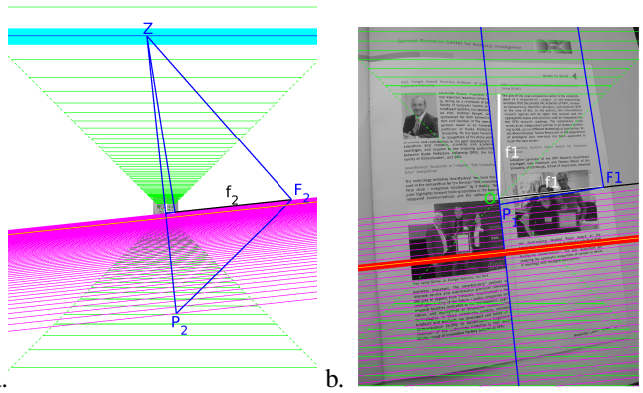


Fig. 4. Zenith and HL search areas and coarse-to-fine sampling. a. Full view. b. Zoom in.

1) Zenith search: The fact that the camera roll is assumed between $-45°$ and $45°$ means that we are looking for the zenith in this angular area relative to the vertical of the image (dashed green lines in Fig. 4). The tilt constraint means that if the focal length f was known, we could limit the zenith search to above $-f$ or below f[1], assuming image coordinates centered at the principal point O, with the y-axis pointing downwards. As we do not know the focal length *a priori*, we use the most permissive focal length in the considered interval, i.e. the smallest one f_1 (Fig. 4.b). The zenith is then estimated by sampling horizontal LSs within the bounded area and then scoring each sample according to Eqn. (2), with a single term in the external sum. The set $\mathcal{A}$ of all LSs detected by LSD is used at this stage. Finally, the zenith Z is first estimated as

[1]Actually, outside the circle of ray f, but we prefer to use the line boundary constraint, which is simpler while leading to same results in our experiments.

the best scored VP along the best scored line sample, and then refined by SVD using the set $\mathcal{V}$ of all LSs consistent with the initial estimate.

2) Horizon line search: Once the zenith is known, we can estimate the zenith line (OZ) (line (P_2Z) in Fig. 4.a) and sample candidate lines perpendicular to the zenith line, according to the general scheme described at the beginning of Sec. II. Actually, if the focal length was known, the HL could be obtained immediately. Indeed, as the HL is in the horizontal plane passing through the optical center F and the zenith is on the world's vertical axis passing through F, line (FP) is perpendicular to the line (ZP), where P is the point of intersection between the zenith line and the HL (Fig. 2). This would allow us to compute the offset $p = ||\overrightarrow{OP}||$ of the HL from $p = f^2/z$, with $z = ||\overrightarrow{OZ}||$. Again, the focal length is unknown at that stage, but this geometrical property allows us to constrain the search of HL in the offset interval $[f_1^2/z, f_2^2/z]$, assuming the offset axis is oriented in the direction opposite to the zenith (Fig. 4.a).

Here the scoring of samples is based on LSs in $\mathcal{A} - \mathcal{V} - \mathcal{H}$, where $\mathcal{H}$ is the set of LSs whose orientation is less than $1.5°$ from that of the normal to the zenith line (i.e. that are almost parallel to the HL). As discussed in Sec. II-B, LSs in $\mathcal{H}$ could generate an infinite VP on each sample, reducing the relevance of the consistency criterion based on the best two scores. Once the offset of the HL is known, we get the focal length $f = \sqrt{zp}$.

B. Coarse-to-fine sampling

In this section, we only describe how HL candidates are sampled within the search area, since the same procedure is used for the zenith. Assuming the camera's tilt angle is uniformly distributed, the HL is first estimated by using a tangent sampling (magenta lines in Fig. 4):

$$\text{offset}(i) = \tilde{f} \tan\left(\theta_{min} + \frac{\theta_{max} - \theta_{min}}{n_1 - 1} i\right), \qquad (3)$$

with $i \in [0, n_1 - 1]$, $\theta_{min} = \tan^{-1} \frac{f_1^2}{z\tilde{f}}$ and $\theta_{max} = \tan^{-1} \frac{f_2^2}{z\tilde{f}}$. Constant $\tilde{f}$ need not necessarily be equal to the true focal length, which is still unknown at this stage. It is preferable, however, that it is of the same order as the image dimensions to avoid too much concentration inside (if $\tilde{f}$ is too small) or outside (if $\tilde{f}$ is too large) the image boundaries. For that purpose, we use $\tilde{f} = \sqrt{WH}$. Once a rough estimate of the HL has been obtained (let's call i_0 the index of the selected sample), we re-sample n_2 samples around this estimate. More precisely, the new samples are given by Eqn. (3), with n_1 replaced by n_2, $\theta_{min} = \left(\theta(\max(0, i_0 - 1)) + \theta(i_0)\right)/2$ and $\theta_{max} = \left(\theta(i_0) + \theta(\min(n_1 - 1, i_0 + 1))\right)/2$, where $\theta(i) = \tan^{-1}(\text{offset}(i)/\tilde{f})$ (red lines in Fig. 4.b).

The numbers of samples n_1 and n_2 are chosen so that $n = n_1 + n_2$ is minimal, while having a dense final sampling between θ_{min} and θ_{max}. Let $\Delta\theta = \theta_{max} - \theta_{min}$, $\beta = \Delta\theta/n_1$ (angle between two samples at the first stage) and $\alpha = \beta/n_2$ (angle between two samples at the second stage). The value of α is set to $\tan^{-1}(1/\tilde{f})$, so that the distance between two samples in the dense area is of the order of a pixel. Setting to 0 the derivative of $n = \frac{\Delta\theta}{\beta} + \frac{\beta}{\alpha}$ with respect to β gives $\beta = \sqrt{\alpha\Delta\theta}, n_1 = n_2 = \sqrt{\frac{\Delta\theta}{\alpha}}$. For example, the total number of samples used for the image in Fig. 4 was 140 (70+70), which is less than half the number of samples used with previous approaches, while a higher sample density is obtained around the predicted offset with our method.

IV. Discretization of Vanishing Points

In this section, we assume that the HL L has been calculated. Horizontal VPs are computed along that line from LSs in $\mathcal{A} - \mathcal{V}$ (LSs in $\mathcal{H}$ are reintegrated in order to be able to detect an infinite VP on that line). Intersecting L with the lines extending the detected LSs should lead to point accumulations around the VPs. A histogram of their x-coordinate (relative to the orthogonal projection P of the principal point O onto line L) should therefore contain peaks at these locations. Unfortunately, detecting these peaks would give imprecise results because of the non-uniformity of the background noise. Indeed, by reducing the image domain to a disc C of centre O and radius 1 and assuming a uniform distribution of the LSs within C, the authors of [22] established that the probability $P(x)$ that a LS intersects L between P and a point X of coordinate x on L (Fig. 5.a) depends on whether L meets C and whether X is inside C or not. The shape of the PDF $\frac{\partial P}{\partial x}(x)$ is shown in Fig. 5.b for different values of $\rho = ||\overrightarrow{OP}||$. Note that it is constant inside the image domain, but decreasing outside. Based on this result, VPs can be discretized through an *a-contrario* reasoning (Sec. IV-B and IV-C), the principle of which is briefly recalled now.

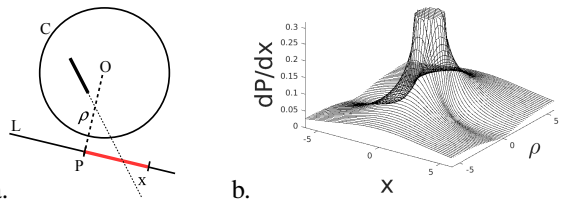


Fig. 5. PDF $\frac{\partial P}{\partial x}(x)$ that a chord of circle C meets line L at coordinate x, with L at distance ρ from the centre of the circle.

A. The a-contrario approach

The *a-contrario* approach is inspired by the principle of Helmholtz, a Gestalt psychologist of the early 20th century, which basically states that "we immediately perceive whatever could not happen by chance". In the mathematical transcription of this principle [23], an event is said to be meaningful if it cannot occur by chance. In our problem, the probability that an extended LS meets line L between coordinates x_1 and x_2 *by chance* is $p = P(x_2) - P(x_1)$. Now let us assume that k extended lines fall in a bin $[x_1, x_2]$ of the x-histogram. The probability that this event occurred by chance is p^k. It is natural to think that, if an event has a high probability of occurring randomly, it is not very meaningful.

However, putting a threshold on the value of p^k would still be somewhat arbitrary. A more founded threshold can be obtained by counting the number of false alarms (NFA), i.e. the number of times this event is expected to occur *by chance* among N bins when M LSs have been detected. This number is given by [23]: NFA $= N\mathcal{B}(M,k,p)$, where $\mathcal{B}(n,k,p) = \sum_{j=k}^{n} \binom{n}{j} p^j (1-p)^{n-j}$ denotes the tail of the binomial distribution of parameters n and p. This calculation can be accelerated by using the large deviation estimate of the binomial tail, based on the relative entropy [23]. An event (here a bin) is said to be ϵ-*meaningful* when its NFA is less than ϵ. When $\epsilon = 1$, the event is simply said *meaningful*. This approach can be extended to a multi-resolution scheme by computing the NFA of all possible combinations of adjacent bins, replacing N by $N(N+1)/2$ in the formula. The authors of [23] call *maximum meaningful mode* (MMM) a set of adjacent bins meaningful in the above sense and maximum in the sense of inclusion of sets of bins.

B. Vanishing point detection

Applying this approach to VP detection, as proposed in [22], raises two problems, both of which are illustrated in Fig. 3.c. Firstly, it is not suitable for curved surfaces because weakly represented VPs can be masked by stronger VPs (in terms of MMMs), which usually results in a low number of VPs being detected on this type of surface. On the other hand, outlier VPs can be obtained due to the accidental meeting of LSs located at distant image locations. We solve these two problems at the same time, by dividing the page into vertical strips, i.e. angular sectors (called *areas* in the following) centred at the zenith in the image plane (see e.g. Fig. 1.a). For each area, we detect one and only one VP, corresponding to the most significant MMM obtained by using only those segments whose leftmost vertex is contained within the area. Thus, a VP obtained in one area cannot be masked by a VP of another area, just as a LS detected in one area cannot be crossed with a LS of another area. However, an important issue remains on the choice of the subdivision to be made. A basic solution would be a fixed number of equal areas, but this would introduce a parameter that cannot be generalized. Indeed, too small areas might not contain enough vanishing LSs, breaking the VP continuity along the HL assumed in the dewarping procedure (Sec. V). On the other hand, too large areas would risk approximating curved areas with flat surfaces, generating an insufficiently fine dewarping in these areas. For this reason, we propose to use a binary-tree descent approach stopping when no MMM is detected in a subregion. This is made possible thanks to the NFA criterion, which allows to decide whether a VP is detected on the basis of a generic threshold ($\epsilon = 1$) obtained through probabilistic reasoning rather than experimental adjustment.

Specifically, our procedure is as follows. First of all, LSs in $\mathcal{A} - \mathcal{V}$ are discretized into pixels and these pixels are projected onto the HL, according to a central projection of center the zenith. A 128-bin x-coordinate histogram of the projected pixels is computed in which empty bins are identified and merged into empty areas if adjacent. If no empty area is detected, a single initial area is obtained, delimited by the lines joining the zenith to the left-most and (resp.) to the right-most projected pixel. If n sets of empty areas are detected, we obtain $n+1$ initial areas according to the same principle, separated by empty areas, as shown in Fig. 6.a where two initial areas are obtained. Each LS in $\mathcal{A} - \mathcal{V}$ is assigned to only one initial area, according to the position of its leftmost vertex. For each initial area, a 128-bin histogram is built from the LSs assigned to the area, and the strongest MMM (i.e. with highest NFA) is detected. As in [22], the histogram is calculated based on bounded values $P(x)$ instead of unbounded x (a uniform PDF must then be used instead of $\partial P/\partial x$). If no MMM is obtained in one or several areas, these areas are added to the set of empty areas. If no MMM is obtained at all, which actually never happened in our experiments, the whole method fails.

Finally, one binary tree root is generated per initial area/MMM pair. The nodes descending from the root are areas, whose angle is halved at each level, paired with MMMs, and the leaves define the final subdivision of the initial area. More precisely, the tree is built by calling the following recursive procedure with a root as input argument:

1) the input node is assigned two children, each one having for area the left or (resp.) right half of the area of its parent and for MMM the one of its parent,
2) for each child: a MMM detection is attempted using the same procedure as for the root with LSs assigned to the child's area. If this fails, the procedure stops for the child who keeps its parent's MMM. If it succeeds, the recursive procedure is called with the child as input argument, whose MMM is replaced by the new one.

This procedure leads to surprisingly fine areas at leaves of the tree. For instance, 417 areas were obtained for the 3024×4032 resolution image shown in Fig. 6.a, whose boundaries are indicated by cyan tics along the HL. Each area gives rise to one VP, initialized from the middle of the highest bin of the leaf's MMM and then refined by SVD using all LSs within the leaf's area consistent with the initial estimate. Small area size coupled with the noise on LSs imply low accuracy of some VPs, which can be corrected through smoothing.

C. Smoothing

VP smoothing is performed in 3D. The empty areas are set aside for smoothing, which is carried out globally on all the leaves of all the trees, sorted in ascending order of the angles between the boundary lines of the areas and the x-axis. The 3D horizontal direction $\overrightarrow{v_i}$ corresponding to a VP (x_i, y_i), i.e. the horizontal direction of the tangent to the surface bounded by the area of the i^{th} leaf, is simply (in the camera coordinate system) $\overrightarrow{v_i} = \overrightarrow{u_i}/||\overrightarrow{u_i}||$ with $\overrightarrow{u_i} = (x_i, y_i, f)$. Smoothing relates to the angles θ_i between these directions and that, on the same plane, of the HL. These angles are between 0 and 180°. In order to avoid discontinuities in the curve around surface strips parallel to the image plane (where the angle can change from a value close to 0° to a value close to 180°),

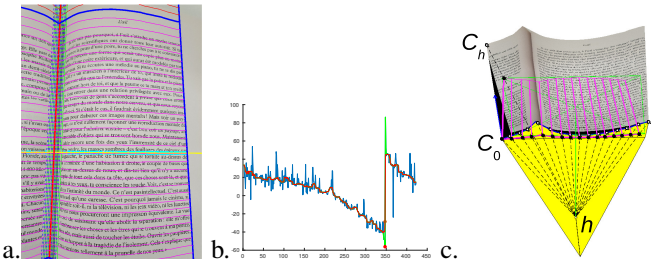


Fig. 6. a. Original image with information on the progress of our algorithm. Empty areas start with a dashed blue line and end with a plain green line (possibly covered by the dashed blue line of the next empty area). Boundaries of the final areas are marked with cyan tics along the HL (in yellow). The breakline is shown in red while the detected borders of the document are drawn in blue. b. VP smoothing. c. 3D reconstruction of the document.

smoothing is performed by removing 180° at angles greater than 90°, resulting in angles in the range [-90°, 90°] (Fig. 6.b, blue curve which corresponds to a path from right to left of the areas).

A median filter is applied to this curve (red curve in Fig. 6.b). In the case of a double-page spread, this smoothing may tend to flatten the "valley" that usually occurs at the junction of the two pages, which corresponds to a discontinuity in the angle of the tangent to the surface (see e.g. on the book in Fig. 6 moving from left to right in the image, the angle changes from about 80° to about -60+180=120°). These discontinuities are detected by simple thresholding of the angle variation after smoothing (red star in Fig. 6.b). If more than one discontinuity is detected, we consider only the largest one. If a discontinuity has been detected, the previous smoothing is replaced by applying a median filter independently on each of the two parts of the curve, before and after the breakpoint (Fig. 6.b, green curve). The breakpoint is converted to a breakline in the image. If the breakpoint was obtained in an area adjacent to an empty area, the breakline is shifted in the middle of the empty area, then divided into two distinct empty areas (Fig. 6.a, red line).

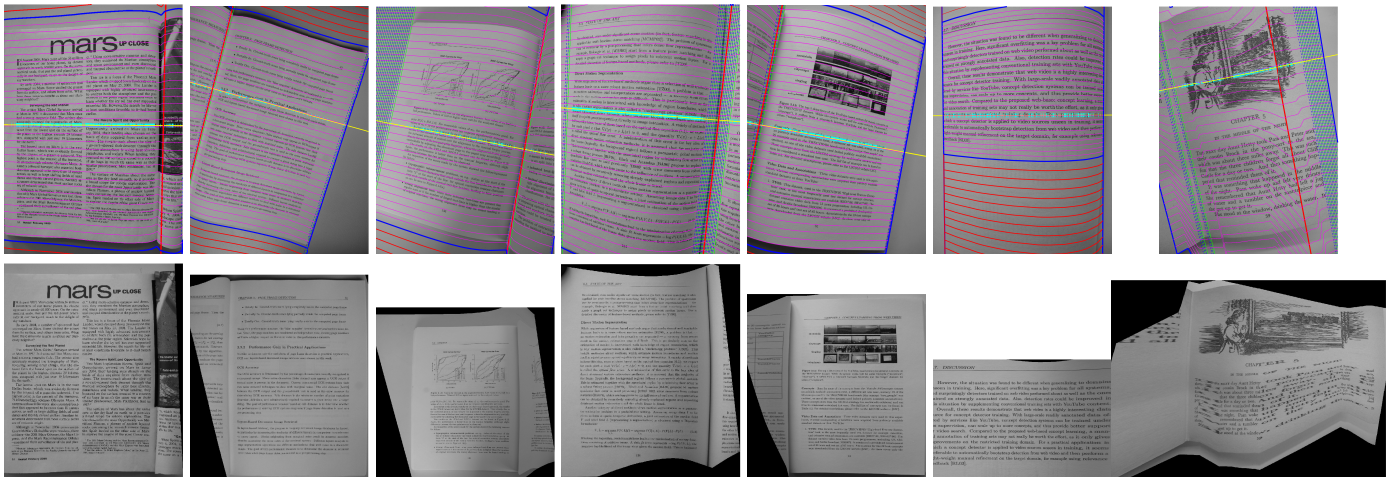
Finally, 3D directions (and therefore VPs) of the empty areas are linearly interpolated from the angles of their left and right neighbours in the general case (e.g. $(\theta_i + \theta_{i+1})/2$ for an empty area between areas i and $i+1$), or extrapolated from their next two (or previous two) neighbours for a peripheral empty area or an empty area adjacent to the breakline (e.g. $2 * \theta_i - \theta_{i+1}$ for an empty area before areas i and $i+1$).

V. Page Reconstruction and Dewarping

Once the focal length and a dense set of VPs are determined in an ordered set of adjacent areas, the 3D reconstruction and then dewarping of the document can be performed easily. We first need to determine the document borders. The left border B_l is taken as the left boundary line of the leftmost area (or possibly as the breakline in case a single, entire page is to be dewarped), and similarly for the right border B_r. The bottom border will start at a point C_0 on B_l. Let's first assume that C_0 is known and consider the area bounded on the left by B_l. The first line segment of the bottom curve connects C_0 to C_1, where C_1 is the intersection point between the right boundary of the considered area and the line passing through C_0 and the area's VP (Fig. 1.a). This process is repeated with the area bounded on the left by the right boundary of the previous area, and C_1 replacing C_0, and so on until the right boundary of the considered area is B_r. The position of C_0 on line B_l is first assumed at the intersection of B_l and the bottom of the image. It is then moved up iteratively with a step of $H/32$ until the generated curve passes above a LS in $\mathcal{A}$. We proceed in the same way to find the top border of the document. Successive curves obtained by this iterative process are drawn in red in Fig. 6.a, and the the final borders in blue. The delimited document is reconstructed in 3D starting from point C_0 to which a z-coordinate equal to f is added. The world's ground plane is defined as the plane passing through C_0 and parallel to the plane containing the HL and the optical centre (yellow polygon in Fig. 6.c). The bottom border of the document is reconstructed by intersecting the inverse rays of the corresponding 2D curve with the ground plane (black and white dots in Fig. 6.c, more distant than in reality for a better visibility). The 3D upper left corner of the document (point C_h in Fig. 6.c) is obtained by intersecting the inverse ray of its 2D counterpart with the world's vertical axis passing through C_0. Finally, the entire document is reconstructed by densely sweeping the bottom curve over the $[C_0 C_h]$ line segment. This generates a dense pixel cloud that just need to be unfolded to get the final dewarped image, which is done by simply butting together the 3D columns of pixels on a 2D plane.

VI. Experimental Results

We estimate our approach on the well known dataset IUPR [25] proposed during the conference ICDAR 2011 for the page dewarping contest. This dataset consists of 100 grayscale document images of pages captured by using hand-held camera. The dataset is defined with different type of layouts (text, images, graphics) large variety of curl with a wide range of perspective distortions and different resolutions. These grayscale images are provided in their binarized version also. As we wrote previously one strength of our approach compared to most of the approaches in the literature is that we need grayscale image only avoiding therefore errors of segmentation due to weak or uneven lighting. In order to objectively compare dewarping methods, our evaluation is focused on character recognition accuracy. The well-known OCR engine Tesseract (version 4.1.1) is used and we compute the word accuracy by counting the number of valid words in the OCR results compared to the GT. Other performance evaluation criteria have been proposed in the literature. In [26] authors propose a performance evaluation methodology but images should be preprocessed to low resolution and some parameters hard to master need to be set. On the other hand using several performance evaluation metrics [27] can lead to misperceptions in performance evaluation. On Tab. I documents captured with a camera are designed as *Distorted* whereas the set *GT scanned* represents the same images but scanned with a flat-bed scanner. *Dewarped* is the set of the



1) 508 / 445 / 508 2) 302 / 0 / 302 3) 244 / 74 / 239 4) 376 / 64 / 370 5) 220 / 2 / 219 6) 180 / 126 / 172 7) 100 / 34 / 97

Fig. 7. Some examples of results obtained on the IUPR dataset. First row: Original images with information on the progress of our algorithm (using the same graphic means as in Fig. 6). Second row: dewarped images. Last row: Number of recognized words for the GT scan / Distorted / Unwarped documents.

	GT scanned	Distorted	Dewarped
Number of words recognized	28608	12481	25865

TABLE I
RECOGNIZED WORDS BY TESSERACT OCR: IUPR DATASET. TOTAL NUMBER OF WORDS ASCII TEXT: 30760

images dewarped using our approach. We can remark from this table the effectiveness of our approach. In percent our method recovers more than 90% of the words compared to the ground truth scanned images. In [19], [18] authors report results around 50% on the same dataset and 34% with the method to compute the VP defined in [28]. Some qualitative results are shown in Fig. 7. Some documents contain images or graphics, others pure text, some are very curved (e.g. the right page in first image), others not so much, and different degrees of camera roll are illustrated. The first line shows the captured images with information about some steps of the algorithm. The second line shows the dewarped images, and the last line the numbers of recognized words, with the same meaning and in the same order as in Tab. I. In order to demonstrate as many results as possible, dewarping is shown on two pages when applicable. However, as GT scanned documents consist of only one page, the OCR was launched on a single page, i.e. the breakline was used to automatically extract the larger area between the breakline and the borders, as for the dewarping results shown in Fig. 8.

The number of words recognized after dewarping is often very close if not equal to the number of words recognized in the GT scanned document, which illustrates the robustness of our method to the various conditions mentioned above. Moreover, our method seems relatively robust to partial non-compliance with the cylindrical model. Indeed, this model is not fully respected at least in images 1,2,4, where the separation line between the two pages is clearly a curved line. This curved shape is also observed in the dewarped images, but the horizontality of the text lines is well respected, leading to a particularly good character recognition in these images. Image 7 shows a case of failure, which comes from the fact that the document contains a drawing with its own perspective. In particular, the drawn window generates a VP that is detected as the strongest over a large area of the document. Thus in the dewarped image, this window is very well orthorectified while the text is particularly distorted. Another part of the document, on the other hand, is correctly dewarped. This constitutes a limitation of our method which may be the object of future work. One approach would be to detect several VPs per area and to search for the smoothest VP path between the left and right borders of the document.

In order to assess the influence of the camera tilt over focal computation and word detection, six 3024×4032 images of the same document page were acquired by a smartphone with tilt varying from near $45°$ to near $0°$ (Fig. 8). The GT focal length was obtained using 9 pictures of a chessboard. The calibration procedure returned a 3233 horizontal / 3227 vertical focal length. The average of these two values is used as GT. Above the distorted images shown in Fig. 8 are displayed the relative error of the computed focal length, and above the dewarped images the number of words detected (the GT is 415). Focal length is the worst calculated in the first and last pictures, and words are the worst detected in the first two pictures. Inaccurate focal length in pictures 1 and 6 are due to the fact that the HL (resp. the zenith) are very far from the image boundaries, although it has little impact on the overall dewarping. On the other hand, the upper part of the document in the first two images appear further away from the camera lens and therefore more blurred than the lower part. As a result, few LSs are detected in the upper part, making dewarping worse in this area, which combined with blurred letters hinders character recognition. This makes several reasons why it is better to avoid such grazing views.

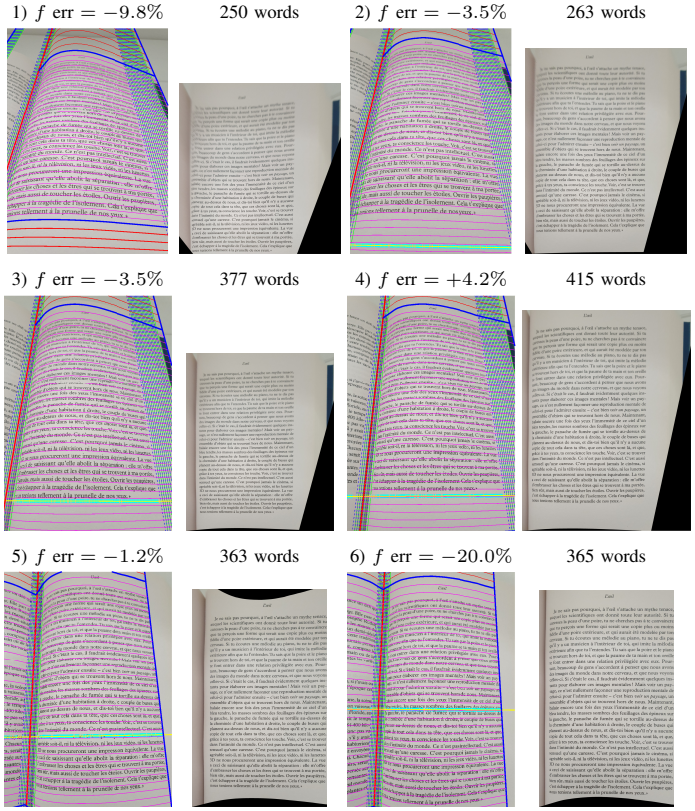


Fig. 8. Influence of the camera tilt on focal computation and word detection.

Fig. 9. An example of result obtained with our method *as it stands* on an image of a building, demonstrating the genericity of our approach. Left: Acquired image with graphic layers showing the intermediate results of the algorithm. Right: Dewarped building.

VII. Conclusion

In this article we propose a new framework which take into account geometric distortions caused by page curl or perspective view. The main advantage of our method is that it can be applied on any kind of document including text, graphics and images. Unlike many other dewarping methods our method does not require any text extraction, line segmentation or image binarization. Moreover, as our approach is generic (Fig. 9), future work will be devoted to urban image 3D reconstruction and dewarping.

References

[1] M. Cutter, M. Michael, and Roberto, "Towards mobile ocr: How to take a good picture of a document without sight," in *15th ACM SIGWEB International Symposium on Document Engineering*, 2015.

[2] A. Ulges, C. H. Lampert, and T. B. Document, "Document capture using stereo vision," in *ACM Symposium on Document Engineering, Milwaukee, Wisconsin, USA*, 2004.

[3] A. Yamashita, A. Kawarago, T. Kaneko, and K. T. Miura, "Shape reconstruction and image restoration for non-flat surfaces of documents with a stereo vision system," in *Int. ICPR*, 2004.

[4] E. Lilienblum and B. Michaelis, "Book scanner dewarping with weak 3d measurements and a simplified surface model," in *14th IAPR International Conference DGCI, LNCS 4992*, 2008.

[5] K. B. Chua, L. Zhang, Y. Zhang, and C. L. Tan, "A fast and stable approach for restoration of warped document images," in *ICDAR*, 2005.

[6] H. Cao, X. Ding, and C. Liu, "A cylindrical surface model to rectify the bound document image," in *ICCV*, 2003.

[7] B. S. Kim, H. I. Koo, and N. I. Cho, "Document dewarping via text-line based optimization," *Pattern Recogntion*, vol. 48, 2015.

[8] T. Kil, W. Seo, H. I. Koo, and N. I. Cho, "Robust document image dewarping method using text-lines and line segments," in *ICDAR*, 2017.

[9] J. Liang, D. DeMenthon, and D. S. Doermann, "Geometric Rectification of Camera-captured Document Images," *IEEE TPAMI*, vol. 30, no. 4, 2008.

[10] Y. Tian and D. G. Narasimhan, "Rectification and 3d reconstruction of curved document images," in *Int. CVPR*, 2011.

[11] A. Ulges, C. H. Lampert, and T. M. Breuel, "Document image dewarping using robust estimation of curled text lines," in *Eighth International Conference on Document Analysis and Recognition (ICDAR)*, 2005.

[12] N. Stamatopoulos, B. Gatos, I. Pratikakis, and S. Perantonis, "A two-step dewarping of camera document images," in *Eighth IAPR International Workshop on Document Analysis Systems, (DAS)*, 2008.

[13] C. Liu, Y. Zhang, and B. Wang, "Restoring camera-captured distorted document images," *IJDAR*, vol. 18, 2015.

[14] H. I. Koo and N. I. Cho, "State estimation in a document image and its application in text block identification and text line extraction," in *European Conference on Computer Vision (ECCV)*, 2010.

[15] K. Ma, z. Shu, X. Bai, J. Wang, and D. Samaras, "Docunet: Document image unwarping via a stacked u-net," 2018.

[16] S. Das, K. Ma, Z. Shu, D. Samaras, and R. Shilkrot, "Dewarpnet: Single-image document unwarping with stacked 3d and 2d regression networks," in *Int. Conf. on Computer Vision (ICCV)*, 2019.

[17] V. K. B. Ramanna, S. Bukhari, and A. Dengel, "Document image dewarping using deep learning," in *7th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, 2018.

[18] Y. Takezawa, M. Hasegawa, and S. Tabbone, "Camera-captured document image perspective distortion correction using vanishing point detection based on radon transform," in *23rd International Conference on Pattern Recognition, ICPR, Cancún, Mexico*, 2016.

[19] ——, "Robust perspective rectification of camera-captured document images," in *7th International Workshop on Camera-Based Document Analysis and Recognition CBDAR@ICDAR, Japan*, 2017.

[20] R. Grompone von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall, "LSD: a Line Segment Detector," *Image Processing On Line*, 2012.

[21] M. Zhai, S. Workman, and N. Jacobs, "Detecting vanishing points using global image context in a non-manhattan world," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[22] G. Simon, A. Fond, and M.-O. Berger, "A-Contrario Horizon-First Vanishing Point Detection Using Second-Order Grouping Laws," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, Sep. 2018.

[23] A. Desolneux, L. Moisan, and J.-M. Morel, *From Gestalt Theory to Image Analysis: A Probabilistic Approach*, 1st ed. Springer Publishing Company, Incorporated, 2007.

[24] Y. Xu, S. Oh, and A. Hoogs, "A minimum error vanishing point detection approach for uncalibrated monocular images of man-made environments," 2013.

[25] S. Bukhari, F. Shafait, and T. Breuel, "The IUPR Dataset of Camera-Captured Document Images," in *Camera-Based Document Analysis and Recognition (CBDAR), LNCS, volume 7139*, 2011.

[26] ——, "An Image Based Performance Evaluation Method for Page Dewarping Algorithms Using SIFT Features," in *Camera-Based Document Analysis and Recognition (CBDAR), LNCS, volume 7139*, 2011.

[27] ——, "Performance evaluation of curled textline segmentation algorithms on CBDAR 2007 dewarping contest dataset," in *ICIP*, 2010.

[28] X. Chen, R. Jia, and H. Zhang, "A new vanishing point detection algorithm based on the hough transform," in *In Third International Joint Conference on Computational Sciences and Optimization*, 2010.